Using Resampling to Assess Reliability of Audio-visual Survey Strategies for Marbled Murrelets at Inland Forest Sites

PATRICK G. R. JODICE[1], STEVEN L. GARMAN[2] AND MICHAEL W. COLLOPY[3,4]

[1]Department of Fisheries and Wildlife, Oregon State University, Corvallis, OR 97331 USA
Internet: jodicep@ucs.orst.edu

[2]Department of Forest Science, Oregon State University, Corvallis, OR 97331 USA

[3]USGS Forest and Rangeland Ecosystem Science Center, 3200 S.W. Jefferson Way, Corvallis, OR 97331 USA

[4]Current address: Department of Environmental and Resource Sciences
University of Nevada, Reno, NV 98512 USA

Abstract.—Marbled Murrelets (*Brachyramphus marmoratus*) are threatened seabirds that nest in coastal old-growth coniferous forests throughout much of their breeding range. Currently, observer-based audio-visual surveys are conducted at inland forest sites during the breeding season primarily to determine nesting distribution and breeding status and are being used to estimate temporal or spatial trends in murrelet detections. Our goal was to assess the feasibility of using audio-visual survey data for such monitoring. We used an intensive field-based survey effort to record daily murrelet detections at seven survey stations in the Oregon Coast Range. We then used computer-aided resampling techniques to assess the effectiveness of twelve survey strategies with varying scheduling and a sampling intensity of 4-14 surveys per breeding season to estimate known means and SDs of murrelet detections. Most survey strategies we tested failed to provide estimates of detection means and SDs that were within ±20% of actual means and SDs. Estimates of daily detections were, however, frequently estimated to within ±50% of field data with sampling efforts of 14 days/breeding season. Additional resampling analyses with statistically generated detection data indicated that the temporal variability in detection data had a great effect on the reliability of the mean and SD estimates calculated from the twelve survey strategies, while the value of the mean had little effect. Effectiveness at estimating multi-year trends in detection data was similarly poor, indicating that audio-visual surveys might be reliably used to estimate annual declines in murrelet detections of the order of 50% per year. *Received 17 January 2001, accepted 29 June 2001.*

Key words.—Accuracy assessment, *Brachyramphus marmoratus*, Marbled Murrelet, Oregon, resampling, survey methodology.

Waterbirds 24(3): 331-344, 2001

Marbled Murrelets (*Brachyramphus marmoratus*) forage in near-shore ocean habitats and typically nest in coastal old-growth coniferous forests throughout much of their breeding range in the Pacific northwest of North America (Gaston and Jones 1998; Nelson 1997). Murrelets are considered threatened outside of their breeding range in Alaska (Kaiser *et al.* 1994; USFWS 1997) and a significant threat to this species has been loss of and disturbance to nesting habitat (FEMAT 1993; Ralph *et al.* 1995). Therefore, there is a need to monitor murrelet populations at forest stands (i.e., nesting habitat) and, to date, observer-based audio-visual (A-V) surveys have been the primary means of doing so.

Audio-visual surveys of Marbled Murrelets were originally designed to determine nesting distribution (Ralph *et al.* 1994). Survey stations are located within or on the edges of forest stands and surveys occur during early morning hours when parents exchange nesting duties. Detections of murrelets (defined as the sighting or hearing of one or more murrelets acting similarly to each other) are tallied for each survey day and, when possible, behavior is recorded to aid in determination of probable nesting status. Typically, four surveys/year are conducted at a stand for two consecutive years. This sampling effort is considered sufficient to determine whether birds are present in an area (the original goal of these surveys; Ralph *et al.* 1994).

There is, however, an unknown correspondence between the number of detections recorded during an A-V survey and either the actual number of birds present in the stand during the survey or the breeding effort in the stand. Furthermore, daily detection data collected from A-V surveys contains substantial temporal and spatial variability, the extent and intricacies of which are obscure (Rodway *et al.* 1993; O'Donnell *et al.* 1995; Jodice and Collopy 2000). Given the need to monitor murrelet numbers and make management decisions for this species at the nest-stand level, counts of daily detections began to be used, both formally and informally, as indices of inter-annual trends in murrelet activity levels and as a means of ranking habitat quality among stands (Miller and Ralph 1995; Stauffer *et al.* 1999). The basic assumption underlying such applications of the detection data was that the number of daily murrelet detections was positively related to either habitat quality or nesting effort. This assumption has yet to be formally tested. Nonetheless, daily detection counts from A-V surveys provide the only data that are consistently available at the nesting stand scale that might be useful for monitoring local numbers of murrelets. Radar surveys of murrelets are being used increasingly, but are not practical in all situations (Burger 1997; Cooper *et al.* 2001).

It is necessary to determine if detection data collected from A-V surveys can be used to accurately monitor murrelet activity and abundance at inland forest stands, especially given the extent of the variability contained within these survey data. While the biological significance of A-V detection data are difficult to address, a purely quantitative assessment of the monitoring value of these same data can be addressed. Our goal was to determine if the magnitude of the intra-stand temporal variability in murrelet detections gathered from A-V surveys was too great to allow managers to use these data to detect trends over time. Our objectives were to: (1) conduct over 50 daily surveys/station throughout the breeding season to obtain empirical estimates of the mean and SD of daily Marbled Murrelet detections; (2) use computer-aided resampling techniques to evaluate how well survey strategies with differing intensity and scheduling estimated within year measures of the mean and SD of murrelet detections obtained during our > 50 daily surveys/station; (3) determine the inter-annual reliability of various survey strategies to detect trends in murrelet detections over time; and (4) expand the evaluation of survey strategies beyond the range of our survey data by producing statistically-generated detection data from an underlying distribution similar to that of the empirical observations. This last objective is particularly critical to ensuring that a full range of murrelet detections (i.e., more than we were able to measure at a limited number of sites during 2-3 years) was considered.

STUDY AREA AND METHODS

Due to the unique nature of the analytical approach taken herein, we provide a list of definitions for key terms. This glossary is located at the conclusion of the methods section.

Study Area

We conducted A-V surveys at seven stations located in five old-growth Douglas Fir (*Pseudotsuga menziesii*) forest stands in the Oregon Coast Range. Each stand was approximately 24 km from the coast and located in the Western Hemlock (*Tsuga heterophylla*) vegetation zone (Franklin and Dyrness 1988) on U.S. Bureau of Land Management lands.

The most northern three survey stations were in the Valley of the Giants region, located near Siletz, Oregon. *Giant 1* (44°56'N, 123°43'W; 365 m a.s.l.) was located along the north fork of the Siletz River while *Giant 2* and *3* (44°55'N, 123°42'W; 535 m a.s.l.) were located in a separate stand on a plateau above the river ca. 2 km from *Giant 1*; *Giant 2* and *3* were approximately 150 m apart. Two survey stations were located along Spencer Creek, a second order tributary of the Umpqua River, about 125 km south of the Valley of the Giants. *Spencer 1* (43°49'N, 123°51'W; 100 m a.s.l.), along the main fork of the creek, and *Spencer 2* (43°49'N, 123°52'W; 100 m a.s.l.), along the upper fork of the creek, were about 1.5 km apart. The most southern two survey stations were located about 90 km south of the Spencer Creek sites along 2x4 Creek, a second order tributary of the Coquille River. Stations *2x4 East* (2x4E) and *2x4 West* (2x4W; 42°52'N, 124°08'W; 425 m a.s.l.) were 500 m apart. With the exception of *Spencer 2*, surveys were conducted at each station prior to our study and results indicated Marbled Murrelets were likely to be nesting within the area covered by each survey station.

Survey Data Collection

The primary sampling unit recorded during surveys and used in our analyses was a "detection". A murrelet detection is defined as "the sighting or hearing of one or more murrelets acting in a similar manner" (Ralph *et al.* 1994; Paton 1995). For example, a flock of flying murrelets sighted during a survey would be recorded as "one detection" with multiple birds. A murrelet vocalization heard during a survey without visual contact would also be recorded as "one detection", although the number of individuals for which the detection referred to would be unknown. Therefore, A-V surveys result in a measure of murrelet activity at a survey station and not in a count of individuals. The Marbled Murrelet Inland Survey Protocol also provides substantial guidance for determining if audio or visual observations of multiple murrelets should count as one or multiple detections (Ralph *et al.* 1994).

All recorded detections were categorized based on observed behavior. For example, some detections recorded murrelet behavior that was more indicative of nesting, such as visual detections of sub-canopy flights or circling flights. These types of detections were referred to as "occupied detections" and we examine this subset of detections separately. Hereafter, we use the term "daily detections" to refer to situations when we pool all detections recorded during a survey regardless of the behavior recorded.

We conducted A-V surveys between 1 May and 4 August 1994 (*Giant 1* and *2*, *Spencer 1* and *2*), 1996 (*Giant 1*, *2*, and *3*), and 1997 (*Giant 1* and *2*, *Spencer 1*, and *2x4E* and *W*) and followed most survey guidelines set forth in the Marbled Murrelet Inland Survey Protocol (Ralph *et al.* 1994). All surveyors were trained to standards set by the Marbled Murrelet Inland Survey Protocol. Surveys began 45 minutes before sunrise and ended 75 minutes after sunrise or 15 minutes after the last detection, whichever was later. Surveys were not conducted during heavy rain or high winds, which would have interfered with visual or aural observations of birds. The same observer surveyed each station during the entire breeding season to eliminate effects of inter-observer variability on within-stand detection data. Two exceptions were at *Giant 1* in 1996 and *Spencer 1* in 1997. Here, we conducted simultaneous A-V surveys for one week with the original and replacement surveyor. Daily tallies of detections and timing of murrelet detections from these simultaneous surveys were similar.

Evaluation of A-V Surveys
to Estimate Detection Means and SDs

We used simulation models to evaluate twelve survey strategies for their ability to produce mean and SD estimates for counts of daily detections and counts of occupied detections that were similar to actual detection mean and SD values we obtained from A-V surveys. The twelve survey strategies we assessed varied in intensity from 4-14 survey days/season, were designed to be logistically feasible, to consider breeding phenology, and, in certain cases, to mimic schedules currently being used by murrelet surveyors (Table 1). Nine of the twelve survey strategies were stratified temporally, while three were random.

We evaluated each of the twelve survey strategies by comparing the mean and SD of the count of daily and occupied detections from the intensive field sampling effort with > 50 survey days/season (hereafter called "observed data") with the mean and SD calculated from the less-intensive simulated surveys that followed the rules of the twelve sampling strategies (hereafter called "samples"). We used Resampling Stats Software (Simon 1995) to simulate survey strategies by randomly selecting sets of survey days to match the rules of the twelve survey strategies. Survey strategies that more closely estimated the observed mean and observed SD were considered to be more reliable. Below is a detailed outline of the procedure that was followed to determine the reliability of each survey strategy at each site and year, and for both occupied and daily detections.

Step 1 selected, without replacement, the appropriate number and temporal distribution of survey days from an observed data set following the rules of a given survey strategy (e.g., select 4 days following the rules of survey strategy protocol 4 [Table 1] from the observed data for the *E2x4* site in 1997). Step 2 calculated the mean and SD for the count of detections for that set of sampled survey days (i.e., sample mean and sample SD). Step 3 calculated the percent difference between the observed mean and sample mean and between the observed SD and the sample SD. Step 4 repeated the first three steps 1,000 times, thus creating 1,000 unique sets of survey days, 1,000 "sampled" detection means and SDs, and 1,000 observed-sample differences. Step 5, the

Table 1. Description of twelve resampling models used to simulate surveys for Marbled Murrelets. Resampling models randomly selected the appropriate number and distribution of days without replacement from audio-visual survey data of Marbled Murrelets collected at seven survey stations in the Oregon Coast Range, 1 May-5 August 1994, 1996 and 1997.

Survey strategy acronym	No. survey days	Temporally stratified (TS) or completely random (CR)	Sampling methods (all days randomly selected)
CR4	4	CR	Selected from entire season
P4[1]	4	TS	1 day from May; 1 day from June; 1 day between 21 June and 21 July; 1 day between 10 July and 4 Aug. At least 6 but no more than 30 days between surveys.
MY4	4	TS	Selected from May
JN4	4	TS	Selected from June
JY4	4	TS	Selected from July
CR7	7	CR	Selected from entire season
BIWK	7	TS	1 day selected from each 2 week period
MY8	8	TS	Selected from May
JN8	8	TS	Selected from June
JY8	8	TS	Selected from July
CR14	14	CR	Selected from entire season
WEEK	14	TS	1 day selected from each week

[1]An approximation of the current Marbled Murrelet survey protocol (Ralph *et al.* 1994).

evaluation stage, calculated the proportion of the 1,000 samples whose mean and SD fell within ±10%, ±20%, and ±50% of the observed mean and observed SD. These three ranges are hereafter called accuracy windows and they are illustrated in Figure 1. We refer to the proportion of samples whose mean and SD fell within each accuracy window as the reliability index for that accuracy window. For example, applying survey strategy protocol 4 to the observed detection data from *E2x4* 1997 resulted in 180 of 1,000 samples having a sample mean and sample SD that were each within ± 20% of the observed mean and observed SD, respectively (Fig. 1, middle box). Therefore, the reliability index for this case was 18%. Survey strategies with higher reliability indices were considered to be more effective at estimating observed means and observed SDs compared to survey strategies with lower reliability indices. These five steps were repeated for each combination of survey strategy (N = 12), site and year (N = 12), and detection type (N = 2); this resulted in 288 sets of resampled surveys, each with 1,000 samples.

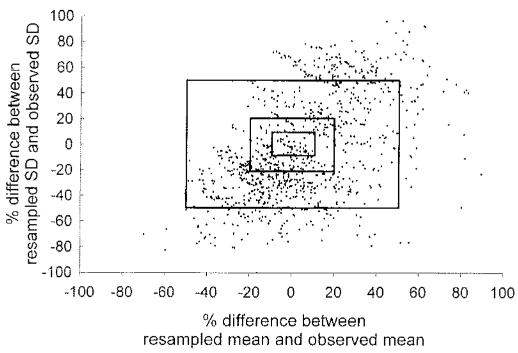

Figure 1. An example of the percentage difference between a known mean and known SD of daily marbled Murrelet detections from field survey data and 1,000 resampled estimates of that mean and SD. Each symbol represents the results of one resampling iteration that followed a rule set that mimics the current Marbled Murrelet Inland Survey Protocol (see P4 in Table 1) and that was compared to detection data gathered at the *E2x4* survey station, 1997. Inner, middle, and outer boxes enclose resampled means and SDs that are within ±10%, ±20%, and ±50% of the observed mean and observed SD, respectively.

Additionally, for each accuracy window and survey strategy, all samples were assigned to one of nine categories based on whether observed means and observed SDs fell below, within, or above the limits of an accuracy window (e.g., Table 2). The proportion of samples within each of the nine categories was calculated for each survey strategy and accuracy window and used to determine the direction and magnitude of error in sample means and sample SDs.

Because the sites we surveyed tended to have relatively low levels of murrelet detections with high variability, the above resampling analysis was restricted to detection data with high SDs and relatively low mean values. We also wanted to determine the reliability of A-V surveys to estimate detection data that had lower levels of variability. Doing so would extend the applicability of our results over a range of murrelet detection levels likely to occur throughout the geographic range of the species. To accomplish this objective, detection data first needed to be computer-generated to complete the range for which we did not have field data. Then, we determined the ability of various A-V survey strategies to estimate these generated detection means and SDs using resampling methodology similar to that described above.

We generated detection data over a range of means that extended from a minimum of 10 detections/day to a maximum of 90 detections/day, and did so in increments of 20 detections/day. For each of these five mean values, we generated data sets with a range of coefficients of variation (CV) extending from 25% to 115% in increments of 10%. This resulted in 50 generated data sets (i.e., one data set for each combination of mean and CV value listed). Of these 50 data sets, 10 had a mean and CV combination that were similar to either our field data or to field data published by Rodway *et al.* (1993). Prior to generating detection data, we evaluated twelve statistical distributions (e.g., normal, Poisson, gamma) to determine which provided the best fit to actual survey data. We chose the gamma distribution, a statistical distribution that is a member of the exponential family of distributions, because it is very flexible in nature, tends to fit count data well, and, most importantly, fitted eleven of our twelve field survey sets well (Kolmogorov Smirnov P > 0.3 for 11 of 12 cases) and also fitted two similarly sized murrelet detection data sets from British Columbia well (Kolmogorov Smirnov P > 0.8; Rodway *et al.* 1993). We also verified the similarity in statistical distributions between generated and observed data that shared similar means and SDs (this amounted to 10 of the 50 data sets) using graphical analyses and by

Table 2. An example of results from a resampling model used to evaluate the reliability of a Marbled Murrelet survey strategy (see Table 1). This example applies survey strategy P4 (Table 1) to murrelet detection data collected at the *E2x4* site in 1997. The value in the cell "mean reliable, SD reliable" is the proportion of 1,000 resampled surveys where the resampled mean and resampled SD were within ±20% of the observed field mean and observed field SD from that site during that year. Values in all other cells are the proportion of the 1,000 resampled surveys that met the definition of the row and column headings. Low = value underestimated (i.e., resampled values less than observed field value by >20%), reliable = value estimated to within ±20%, and high = value overestimated. Identical matrices were generated for each site (N = 7), year (N = 3), survey strategy (N = 12), and accuracy window (N = 3) combination.

	Mean low	Mean reliable	Mean high
SD low	12.3	22.7	4.8
SD reliable	6.3	18.0	6.6
SD high	0.1	13.6	15.6

comparing results from resampling exercises using the completely random survey strategies previously described (see Jodice 1999 for details). Details of the function used to generate the gamma variates can be found in Evans *et al.* (1993). Below we provide the details of the data generation and resampling procedure.

The data generation and resampling process included four steps and each step was conducted for all 50 data sets described above; to clarify the explanation of the process, however, we review the steps using as an example a data set with mean = 10 and CV = 25%. Step 1 generated three sets of variables with sample sizes of 4, 7, and 14. These sample sizes simulated completely random surveys for daily detections with 4, 7, and 14 days of effort at a site with known mean (e.g., 10) and SD (e.g., 2.5). Each set of variates was drawn from a distribution that had a mean = 10 and SD = 2.5 which was generated with the SAS procedure RANGAM (SAS Institute, Inc., 1985). Step 2 calculated the mean and SD of each generated data set created in Step 1. This mean and SD are equivalent to the sample mean and sample SD described previously. Step 3 calculated the percent difference between the known mean used to generate the distribution from which the samples were drawn and the sample mean, and between the known SD and the sample SD from each of the three generated data sets. Step 4 repeated the first three steps 1,000 times. Step 5 calculated reliability as the proportion of the 1,000 generated samples (from each sample size) whose mean and SD were within ±10%, ±20%, or ±50% of the known mean and known SD from the cell definition. These steps were repeated for each of the generated data sets (N = 50) and for each of the three sample sizes (N = 150); this resulted in 7,500 sets of generated surveys, each with 1,000 samples.

Evaluation of A-V Surveys to Detect Inter-annual Trends

While the above procedures focused on determining reliability of survey strategies to estimate detection means and SDs within one survey season, the following procedure focused on determining reliability of survey strategies to estimate declines in detection means over multiple survey seasons. To accomplish this, we generated a series of data sets to represent multiple years of survey effort and built into each a known decline in the mean (i.e., to simulate a decline in the mean value of daily detections). We also examined the effect of intra-annual variability in detections on survey reliability by varying the CV of the generated data sets. For ease of discussion, we use the terms for the elements that were being simulated (i.e., decline in detections, survey effort/year, and years of surveys). We completed these analyses for all combinations of two rates of decline (25% and 50% per year), three levels of intra-annual survey intensity (4, 7, and 14 survey days scheduled at random), two levels of inter-annual survey intensity (3 and 5 years of surveys), and two levels of intra-annual variability in detections (45% and 85%). We outline the steps below with an example of a simulation that uses four randomly scheduled surveys per year for three survey years to estimate a mean that is declining by 25% per year and which has an intra-annual CV of 45%.

Step 1 used SAS procedure RANGAM (SAS Institute, Inc., 1985) to generate a sample of four detections from a gamma distribution with a mean of 50 and CV of 45%. The sample sizes of four represented a survey with four days of effort scheduled at random. Step 2 generated two additional sets of four detections from a gamma distribution where the original mean of 50 was reduced by 25% per year (i.e., per data set) but where the CV remained constant at 45%. Steps 1 and 2 represented, therefore, three survey years during which time the mean number of murrelet detections decreased by 25% per year. Step 3 calculated the slope of sampled detections over time by regressing the simulated data sets upon year number. Step 4 calculated the difference between the slope that was built into the gamma distributions from which the detections were drawn (i.e., -25% per year) and the resampled slope from the simulated survey. Step 5 repeated steps 1-4 a thousand times and then calculated the proportion of the 1,000 samples whose slopes were within ±10%, ±20%, ±30%, ±40%, or ±50% of the actual slope. This proportion is referred to as the inter-annual reliability index. These steps were repeated for each of the eight possible combinations of annual rate of detection decline (-25% per year and -50% per year), number of survey years (three or five year periods), survey effort per year (4, 7, or 14), and level of variability in detections (CV = 45% or 85%); this resulted in 24 sets of simulations, each with 1,000 samples.

Glossary of Terms

Detection: the sighting or hearing of one or more Marbled Murrelets. If more than one bird was involved, they acted similarly to each other.

Daily detections: the daily tally of all murrelet detections recorded during an A-V survey.

Occupied detections: a subset of daily detections where behavior that was more indicative of murrelet nesting was recorded.

Observed data: the mean and SD of the count of daily or occupied murrelet detections from the intensive field sampling effort.

Samples: the mean and SD calculated from simulated surveys conducted via resampling that followed the rules of the sampling strategies noted in Table 1.

Reliability: a measure of the effectiveness of a survey strategy. Survey strategies that produced sample means and sample SDs that more closely estimated the observed mean and observed SD were considered to be more reliable.

Reliability index: a means of quantifying the above concept. The reliability index is the proportion of 1,000 samples whose mean and SD fell within a predefined limit of the observed mean and observed SD.

Accuracy window: A predefined limit within which sample means and sample SDs must fall to be considered "reliable". Three such windows were used: ±10, ±20, and ±50% (see Fig. 1).

Gamma distribution: a statistical distribution that is a member of the exponential family of distributions and which is very flexible in nature and tends to fit count data well (see Evans *et al.* 1993 for equations used to generate gamma distributions).

Generated data sets: A set of variables of sample size 4, 7, or 14 that were drawn from a gamma distribution with known mean and known SD and used to represent a survey effort for murrelet detections of 4, 7, or 14 days scheduled at random. Generated data sets may simulate one or multiple years of survey effort.

Inter-annual reliability: a measure of the effectiveness of a survey strategy to estimate a decline in a mean value of generated data that represent murrelet detections. Survey strategies that produced sample slopes that more closely estimated the known slope were considered to be more reliable.

Inter-annual reliability index: a means of quantifying the above concept. The inter-annual reliability index is the proportion of 1,000 samples whose slopes fell within a predefined limit of the observed slope. The limits used were ±10, ±20, ±30, ±40, and ±50%.

Means are presented ± 1 SD unless noted otherwise.

RESULTS

We conducted 681 Marbled Murrelet surveys, averaging 56.7 ± 4.5 survey-days at each survey station during each breeding season. We recorded 16,105 detections and 33.5% of these were classified as "occupied detections". One or more detections were recorded in 90% of surveys, although nine of twelve site-by-year combinations had at least one day without detection. We failed to record any occupied detections during 26% of the surveys. Seasonal means of daily detections ranged from 7.7 to 51.3 detections/day among all sites and years, while seasonal means of occupied detections ranged from 1.2 to 27.6 detections/day (Table 3). The CV for daily and occupied detections within each site and year also varied greatly (Table 3). The grand mean of CVs from occupied detections among all sites and years (134 ± SD 41.5) was, in fact, greater than that for daily detections (97.5 ± SD 32.0; paired t_{11} = 3.7, $P < 0.005$). It was not uncommon to observe near-minimum and near-maximum counts of daily and occupied detections at a station within the same week (e.g., Fig. 2). Intra-annual variation in counts of daily detections within stations was not strongly or consistently related to date or weather (Jodice and Collopy 2000). There was substantial inter-annual variability in means of daily detections within sites between years (Table 3). Additional details of daily detection data are available in Jodice (1999) and Jodice and Collopy (2000).

Reliability of Survey Strategies

Most of the survey strategies we evaluated did not provide reliable estimates of daily detection means or detection SDs (Fig. 3). On average, <15% and <40% of resampled surveys from any survey strategy provided estimates of the mean and SD of daily detections that were within ±10% or ±20% of observed values, respectively (Fig. 3a-b). When accuracy criteria were relaxed to ±50%, two survey strategies (completely random 7 and one survey conducted biweekly) resulted in average reliability indices near 60% (Fig. 3c) and two survey strategies (completely random 14 and one survey conducted weekly) resulted in average reliability indices near 80% (Fig. 3c).

Table 3. Mean counts and coefficient of variation of daily Marbled Murrelet detections from seven survey stations in the Oregon Coast Range, 1 May-5 August 1994, 1996, and 1997. Occupied detections are a subset of all detections that are considered to be more indicative of nesting. N = number of survey days.

			All detections		Occupied detections	
Site	Year	N	Mean	CV	Mean	CV
Spencer 1	1994	63	32.4	130	27.6	134
Spencer 2	1994	58	16.2	134	4.2	133
Giant 1	1994	55	27.3	68	7.4	115
Giant 2	1994	55	36.1	49	6.5	75
Giant 1	1996	48	7.7	113	1.3	223
Giant 2	1996	51	14.1	72	2.3	126
Giant 3	1996	51	16.2	83	1.2	133
Spencer 1	1997	59	10.6	152	6.8	141
Giant 1	1997	57	15.3	87	2.9	138
Giant 2	1997	58	14.7	122	4.1	198
E2x4	1997	63	51.3	69	17.3	84
W2x4	1997	63	34.2	88	8.7	111

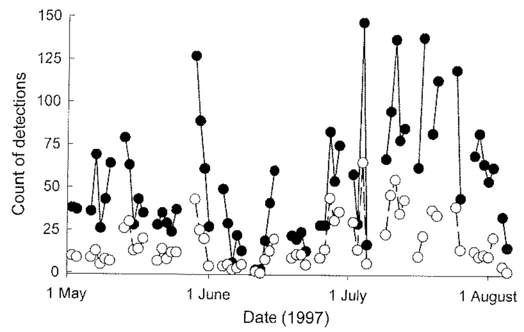

Figure 2. Daily counts of Marbled Murrelet detections at the *E2x4* survey station, Oregon Coast Range, 1 May-4 August 1997. Closed circles with solid lines are all detections, while open circles with dashed lines are the subset of occupied detections.

Similar results occurred when occupied detections were examined separately (Fig. 4). Each survey strategy we evaluated estimated the mean and SD of the daily count of occupied detections to within ±10% or ±20% of observed data in fewer than 30% of cases, on average (Fig. 4 a-b). When accuracy criteria were relaxed to ±50%, two survey strategies (completely random 7 and one survey conducted biweekly) resulted in average reliability indices near 70% (Fig. 4c) and two survey strategies (completely random 14 and one survey conducted weekly) resulted in average reliability indices near 90% (Fig. 4c). Resampled surveys were most likely to simultaneously underestimate both the mean and SD of daily and occupied detections in accuracy windows ±10% and ±20% and underestimate the SD but reliably estimate the mean in accuracy window ±50% (Jodice 1999).

Temporally stratified surveys did not provide more reliable estimates of detection means or detection SDs compared to completely random surveys. For example, there was no significant difference in reliability between temporally stratified and completely random surveys with four days of effort within the ±10% accuracy window ($F_{4,51}$ = 1.82, n.s.; Fig. 3a). The only significant differences ($F_{4,51} > 3.10$, $P < 0.02$) within accuracy windows ±20% and ±50% indicated that the survey strategy with 4 survey days all selected in May was least reliable (Fig. 3b-c). Among survey strategies with 7 or 8 days of effort, there also was no significant difference ($F_{4,51} < 2.5$, n.s.) in reliability within the ±10% or ±20% accuracy window between temporally stratified and completely random surveys (Fig. 3a and b). There was, however, a significant difference between temporally stratified and completely random surveys for seven or eight day surveys within accuracy window ±50% ($F_{4,51} > 4.33$, $P < 0.004$; Fig. 3c) where the survey strategy with 8 survey days, all from May, was the least reliable. There were no significant differences in reliability within any accuracy window between survey strategies comprised of 14 survey days selected completely at random versus one survey day selected each week for 14 weeks (t_{11} < 1.62, n.s.; Figs. 3a-c). Similar results were observed when occupied detections were considered separately; in no situations (i.e.,

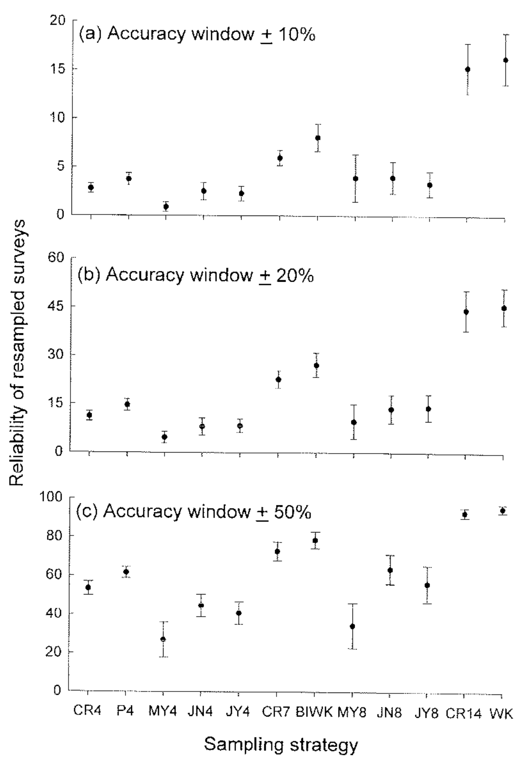

Figure 3. Reliability of resampled surveys to provide estimates of daily Marbled Murrelet detection means and SDs that were within (a) ±10%, (b) ±20%, and (c) ±50% of actual murrelet daily detection means and SDs as estimated from intensive field surveys at seven stations in the Oregon Coast Range, 1994, 1996, and 1997. Values are mean (±1 SE) reliability across all sites and years (N = 12). See Methods for definition of reliability. Note that the scale of the y-axis changes among plots. Rules and acronyms of survey strategies are presented in Table 1.

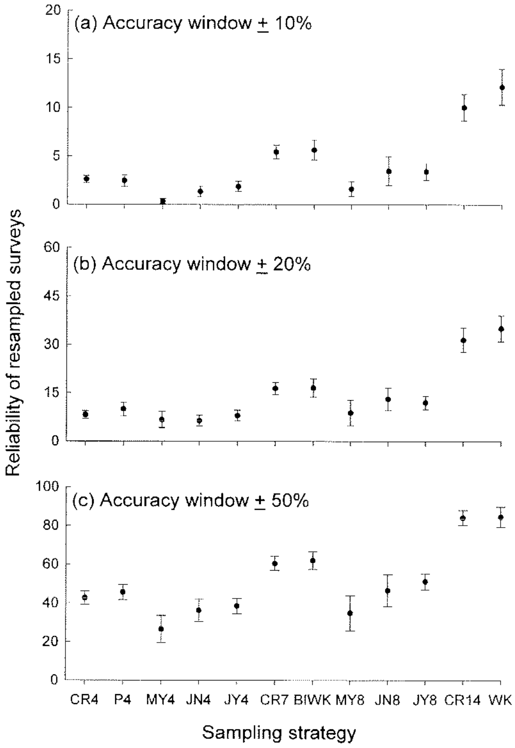

Figure 4. Reliability of resampled surveys to provide estimates of daily Marbled Murrelet occupied detection means and SDs that were within (a) ±10%, (b) ±20%, and (c) ±50% of actual murrelet daily occupied detection means and SDs as estimated from intensive field surveys at seven stations in the Oregon Coast Range, 1994, 1996, and 1997. Values are mean (±1 SE) reliability across all sites and years (N = 12). See Methods for definition of reliability. Note that the scale of the y-axis changes among plots. Rules and acronyms of survey strategies are presented in Table 1.

across all accuracy windows and for each category of survey effort) did temporally stratified surveys outperform completely random surveys (Figs. 4a-c).

Because temporally stratified surveys never outperformed random surveys, we chose to use only completely random survey strategies of 4, 7, and 14 days for the analyses of survey reliability with the broader range of generated data, and for analyses of interannual reliability with multiple years of generated detection data. Furthermore, we based these analyses on characteristics of the complete set of counts of daily detections because we observed little to no difference in the reliability with which either completely random or temporally stratified surveys estimated daily detections versus occupied detections. Nonetheless, the results from the following analyses are as applicable to daily detections as they are to occupied detections (see Discussion).

Survey reliability improved as the variability within the generated data sets declined. For example, with only four days of survey effort and a CV of only 45% (which is slightly lower than the minimum CV of 50% we recorded within a survey year; Table 3), nearly 70% of resampled surveys estimated both the mean and SD of the generated data sets to within ±50% of actual values (Fig. 5a upper tier of data points). Similarly, reliability approached 90% with seven days of survey effort when the CV was 45% and the accuracy window was set at ±50% (Fig. 5b, upper tier of data points). Increasing the sample size to 14

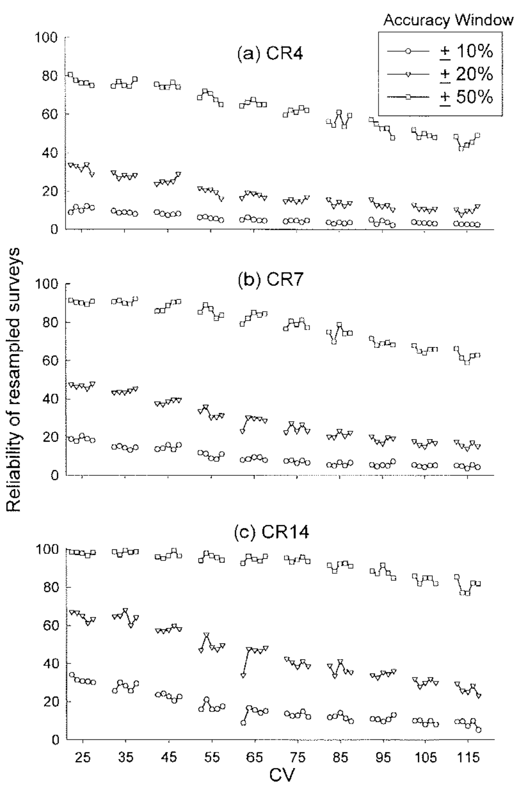

Figure 5. Reliability of resampling routines (N = 1,000 iterations) with a completely random sampling effort of (a) 4, (b) 7, or (c) 14 selections to provide estimates of known means of 10, 30, 50, 70, or 90 (represented by each quintet of data points) and CVs ranging from 25% to 115%. Known means and SDs were generated from a series of gamma distributions.

resulted in a reliability of 100% when the accuracy window was set to ±50% and the CV was 45%. However, surveys proved to be less reliable if a more accurate estimate was desired. For example, within the ±20% accuracy window, a reliability index of 70% could only be achieved when the CV for detections was reduced to 25% and the sample size set at 14 days (Fig. 5a-c middle tier of data points in each plot). Survey strategies with 4, 7, and 14 survey days rarely produced a sample mean and sample SD that were within ±10% of the known mean and known SD (Fig. 5a-c, lowest tier of data points in each plot). Consistent differences or patterns in reliability values were not apparent among means within or among CV values (Fig. 5).

Reliability of Survey Strategies to Detect Annual Trends in Detections

The accuracy of estimating 25% and 50% annual declines in detection means with regression analyses was strongly affected by the intra-annual CV (i.e., 45% or 85%), the annual survey effort, and the number of years over which surveys were conducted (Fig. 6). The inter-annual reliability of most survey scenarios to estimate a decline of 25% per year to within ±30% of the actual decline was <70% (Fig. 6a, c, d, lines with solid symbols). For example, one of the optimum scenarios for accurately estimating a decline in detections of 25% per year required five years of surveys, seven surveys per year, and an intra-annual CV in detection counts of 45%. The inter-annual reliability of this survey effort to estimate the stated decline to within ±30% was near 80% (Fig. 6b). When attempting to estimate a 25% per year decline with a CV of 85% instead of 45%, however, inter-annual reliability surpassed 70% only when 14 surveys/year were conducted for five years, and when the accuracy to which the stated decline was estimated was as poor as ±40% (Fig. 6d).

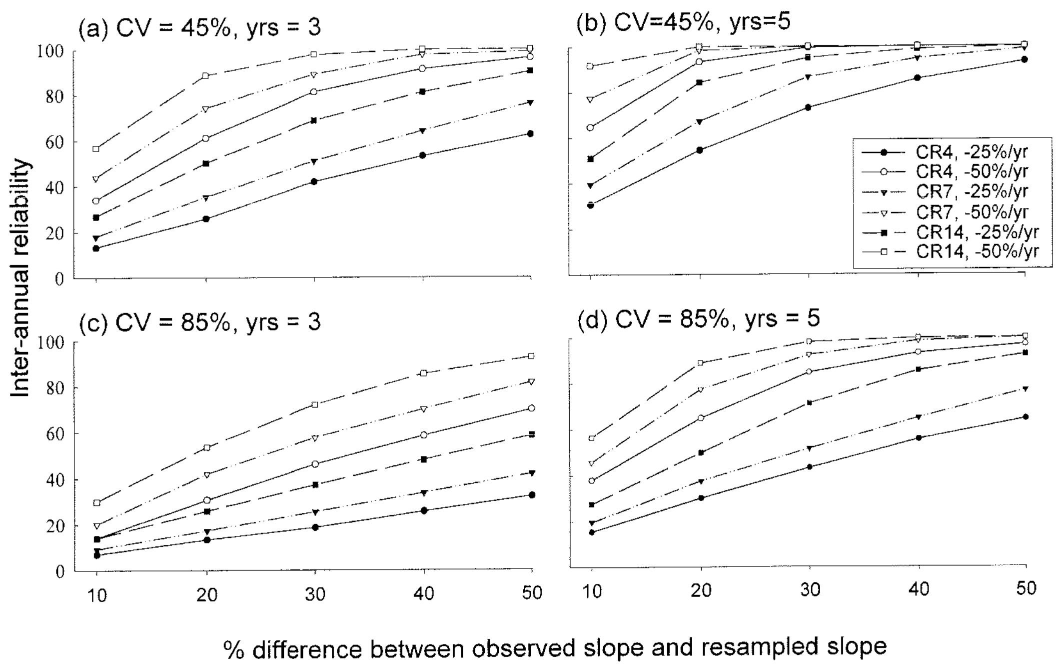

Figure 6. Inter-annual reliability of resampling routines (1,000 iterations) with completely random sampling effort of 4, 7 or 14 selections to estimate known annual declines of 25% and 50% per year to within ±10%, ±20%, ±30%, ±40%, or ±50% of a known slope. Data were generated from a series of gamma distributions. Inter-annual reliability = the proportion of the 1,000 iterations where the estimated slope from resampling routines was within one of the five aforementioned accuracy ranges: (a) within-year CV of generated data = 45%, years of surveys = 3; (b) CV = 45%, years of survey = 5; (c) CV = 85%, years of survey = 3, and; (d) CV = 85%, years of survey = 5.

Annual declines in detections of 50% per year were more reliably estimated (Fig. 6a-d, open symbols). For example, when the CV was 45% and only four surveys/year were conducted for three years, inter-annual reliability surpassed 70% at an accuracy level of ±30% (Fig. 6a). Conducting two additional years of surveys substantially increased both the inter-annual reliability of the surveys and resulted in more accurate estimates of the decline (Fig. 6b). In contrast, when the CV was 85% and only four surveys/year were conducted for three years, regression analyses were only able to estimate the magnitude of the decline with an inter-annual reliability ≥70 to within ±50% (Fig. 6c). With a CV of 85% and three years of surveys, estimating the 50% annual decline to within ±30% was only possible when 14 surveys were conducted/year (Fig. 6c). Conducting two additional years of surveys substantially increased both the inter-annual reliability of the surveys and resulted in more accurate estimates of the decline. For example, four surveys/year estimated the decline to within ±30% with an inter-annual reliability of ≥70 (Fig. 6d).

DISCUSSION

Implications for Monitoring Marbled Murrelets

Neither temporally stratified or completely random survey efforts with 4-14 survey days/season regularly provided estimates of the mean and SD of daily Marbled Murrelet detections to within ±10% or ±20% of observed values. However, the survey strategies we evaluated regularly provide estimates of daily detections to within ±50% of observed values when survey effort was seven or more days/season. Survey efforts of four days/season may prove reliable for wide accuracy windows (i.e., ±50%) if SDs are low (i.e., <50%), although temporal variability in daily murrelet detections tends to be higher than this (Rodway *et al.* 1993; Jodice and Collopy 2000). The ability to estimate the mean and SD of daily detections to within ±50% at a site during a single breeding season suggests that differences in the annual mean number of detections on the order of 50% should be detectable. This should allow murrelet surveyors to detect changes in numbers of daily detections of catastrophic or extreme proportions, but not of smaller changes.

The survey strategies we evaluated also failed to provide consistently reliable estimates for the daily mean and SD of occupied detections except when accuracy criteria were low (i.e., ±50%) and survey effort was seven or more surveys/season. This is not surprising given that the temporal variation in occupied detections was higher than the temporal variation in daily detections at our sites. It is important to recognize that the sites we surveyed had relatively low counts of occupied detections with high temporal variability. This was unlikely to be due to our inability to observe occupied detections because of limited visibility, but rather to murrelet activity patterns that infrequently included "occupied" behavior. Therefore, it would be valuable to conduct analyses similar to ours at stands with a greater frequency of occupied detections so as to better document the temporal variability and seasonality of those particular detections and to determine if seasonally stratified survey efforts might provide more reliable estimates of occupied detections than temporally random surveys. Nonetheless, our analyses with generated data clearly demonstrated that survey reliability for count data increased when temporal variability in the survey target decreased. Therefore, if occupied detections are found to be less variable than daily detections at other locations (e.g., perhaps where nesting density was greater), then daily counts of occupied detections might provide a more reliable monitoring metric.

Temporal variability in daily Marbled Murrelet detections was higher in this study than in the only other study with a similar level of survey effort (Rodway *et al.* 1993). Temporal variability in occupied detections have not been reported from studies with survey efforts similar to ours, so it is difficult to determine how temporal variability in occupied detections from our surveys compared to that at other locations. Nonetheless, high levels of daily variability in each detection count contributed strongly to the low reliability provided by most of our survey strategies. This is supported by the inverse relationship between SD and reliability with generated data. Furthermore, the large day-to-day variability in detections within survey stations was the likely reason completely random survey strategies usually provided comparable reliability indices to temporally stratified surveys within each accuracy window. Preliminary explorations of ten other temporally stratified and two other completely random survey strategies, each with 5-15 days of survey effort, similarly yielded low levels of reliability (P. G. R. Jodice, unpublished data). Therefore, the survey strategies we tested did not adequately account for the high levels of daily or annual variability in detections and did not consistently provide reliable estimates of observed detection means and SDs.

Interpretation of survey results can be improved by examining the type, magnitude, and direction of errors observed from resampled surveys. For example, most surveys usually underestimated both the mean and SD of detection counts in the ±10% and ±20% accuracy windows. It is likely that this occurred because of the prevalence of days where we recorded few or no detections; this was especially true for the subset of occupied detections. Therefore, researchers using either type of murrelet detection data could expect that estimates of the SD from survey efforts of 4-7 days would be underestimates. This is particularly important if temporal or spatial differences in detections were being sought, as such analyses would be more likely to result in a statistically significant difference if SDs were underestimated rather than overestimated. In such a situation sample sizes should be increased to provide more reliable estimates of both the mean and SD. In contrast, underestimates of means may not be as derisive to monitoring or research as long as the bias in the count was consistent; i.e., patterns could still be detected with artificially low numbers although the actual magnitude of the counts might be in error.

Given that single-year results indicated that reliability of most survey strategies was poor when attempting to estimate means and SD to better than ±50%, it was not surprising that the inter-annual reliability of multi-year regression analyses to estimate annual declines in detections (i.e., generated means) of 25% was low in most of the survey scenarios we assessed. Regression analyses using generated data clearly showed that as the CV increased (e.g., Fig. 6a versus 6c) the inter-annual reliability to estimate the trend in the mean decreased substantially. Nonetheless, an increase in survey effort from three to five years with a CV of either 45 or 85% sufficiently increased inter-annual reliability so that substantially fewer surveys within each year were required. These results should be interpreted conservatively, however, as any increase in CV from one year to the next would decrease inter-annual reliability and accuracy.

Survey Effort Recommendations

Our results clearly showed that murrelet detection data recorded during audio-visual surveys had high levels of within-site variability that distorted estimates of daily detection means and SDs. Nonetheless, detection data may, at times, provide the only available data for monitoring or decision-making purposes. Therefore, we provide the following guidelines to assist in survey design. For the purposes of this discussion, we assume that the intra-annual CV for counts of detections is between 50% and 100%.

- Four surveys/season will provide a reliable estimate of detection means and SDs to within ± 50% of actual values in approximately 60% of cases if the Protocol schedule (Table 1) is used or if a completely random schedule is used.
- Seven surveys/season will provide a reliable estimate of detection means and SDs to within ±50% of actual values in approximately 70% of cases if either a completely random schedule is used or if sampling occurs once in every two-week period.
- Fourteen surveys/season will provide a reliable estimate of detection means and SDs to within ±50% of actual val-

ues in approximately 90% of cases or to within ±20% of actual values in approximately 50% of cases, if either a completely random schedule is used or if sampling occurs once every week.

- Sample sizes should be increased to ≥14 surveys/season whenever possible. The greatest benefit of doing so is the improved accuracy with which the SD of detection counts is estimated.
- The majority of surveys within a year should not be conducted during May.
- If the intra-annual CV for detection counts ≤45%, then declines in detection counts on the order of 50% per year may be estimated to within ±20% with seven surveys/year for three years, or with four surveys/year for five years.
- If the intra-annual CV for detection counts is approximately 85%, then it may be most reasonable to seek a 50% per year decline over five years with seven surveys/year or over three years with 14 surveys/year.

Implications for Use with Count Data and Management Recommendations

Evaluating the performance of survey strategies as we described herein is not unique, although examples in the published literature are few (Schwagmeyer and Mock 1997). Such analyses may be viewed as an extension of pilot studies. For example, preliminary data may be used to test the effectiveness of various sampling strategies and subsequently justify the selected sampling strategy, a rarely documented decision (Beier and Cunningham 1996). Such analyses also may provide useful estimates of variance for the metric of interest that can then be used in retrospective and prospective power analyses (Ribic and Ganio 1996; Steidl *et al.* 1997). Additionally, analyses such as ours have the advantage of being applicable to other metrics of central tendency, such as the median. Furthermore, it would not be difficult to apply this analytical process to count data from statistical distributions other than the gamma, given that they provided a good fit to the data of interest (Beier and Cunningham 1996).

Our results can also be applied to analyses of detection data with regard to spatial issues, given that variability due to interobserver variability and site-specific environmental differences could be accounted for. Therefore, using detection data to compare habitat quality among stands may be prone to the same magnitudes of error as using detection data to seek annual trends in activity over time within stands. Our analyses do not suggest the need to make any changes to the current Marbled Murrelet survey protocol, particularly with respect to determining the optimum survey effort necessary for detecting occupancy at a stand. This is because such an analysis would require an abundance of survey data that was collected at individual stands during two consecutive years; we only obtained survey data from the same site for two consecutive years at two survey stations (*Giant 1* and *2*) and this was insufficient for conducting such an analysis.

In conclusion, our data clearly showed that the level of variability in count data must be assessed prior to developing long-term monitoring plans. Regardless of the unit being counted, even moderate levels of variability can require substantial sampling effort to overcome the poor performance of most survey strategies. The fact that many of our survey efforts underestimated temporal levels of variability is particularly disconcerting, as this may increase the likelihood of finding statistically significant differences when in fact none may exist. Additional, site-intensive surveys of murrelet detections (both daily and occupied), particularly at sites with higher levels of activity and nesting, would provide needed detail on consistency in seasonal detection patterns and temporal variability among sites. We also suggest that similar, long-term data sets be collected from other portions of the species range to document the degree of geographic variability in detection patterns. Based on the results of our analyses, we suggest that the use of Marbled Murrelet detection data for quantitative analyses be limited and the appropriateness considered at length prior to initializing research, management, or monitoring efforts.

ACKNOWLEDGMENTS

The Oregon State Office of the Bureau of Land Management, the USGS Biological Resources Division, and the USGS Forest and Rangeland Ecosystem Science Center funded this research. We are grateful to the following field technicians who endured many rainy mornings collecting survey data: Betsy Arden, Tracey Arensberg, Brad Buckley, Kemper Carlsen, Kelley Goocher, Lyndia Hammer, Rick Hatcher, Barry Hughes, Chris Knauf, Molly Monroe, Dana Morley, Mike Reardon, Rob Schneider, and Katerine Schletz. We also thank BLM biologists Joe Lint, Barb Hill, Wayne Logan, Larry Mangan, Kevin Kritz, and Gary Likata for their support, advice, and assistance with field logistics. Lisa Ganio and Manuela Huso provided statistical support. Alan Burger, John Coulson, Robert Jarvis, Laura Jodice, and Fred Ramsey provided constructive reviews of the manuscript.

LITERATURE CITED

Beier, P. and S. C. Cunningham. 1996. Power of track surveys to detect changes in cougar populations. Wildlife Society Bulletin 24: 540-546.

Burger, A. E. 1997. Behavior and numbers of Marbled Murrelets measured with radar. Journal of Field Ornithology 68: 208-223.

Carter, H. R. and Kuletz, K. J. 1995. Mortality of Marbled Murrelets due to oil pollution in North America. Pages 261-270 *in* Ecology and conservation of the Marbled Murrelet (C. J. Ralph, Jr., G. L. Hunt, M. G. Raphael and J. F. Piatt, Eds.). Pacific Southwest Research Station Forest Service, USDA General Technical Report PSW-GTR-152. Albany.

Carter, H. R., M. L. C. McAllister and M. E. Isleib. 1995. Mortality of Marbled Murrelets in gill nets in North America. Pages 271-284 *in* Ecology and conservation of the Marbled Murrelet (C. J. Ralph, Jr., G. L. Hunt, M. G. Raphael and J. F. Piatt, Eds.). Pacific Southwest Research Station Forest Service, USDA General Technical Report PSW-GTR-152. Albany.

Cooper, B. A., M. G. Raphael and D. Evans-Mack. 2001. Radar-based monitoring of Marbled Murrelets. Condor 103: 219-229.

Evans, M., N. Hastings, and B. Peacock. 1993. Statistical distributions. John Wiley and Sons. New York.

Forest Ecosystem Management Assessment Team. 1993. Forest Ecosystem Management: An ecological, economic, and social assessment. USDA Forest Service Pacific Northwest Region, Portland.

Franklin, J. F. and C. T. Dyrness. 1988. Natural vegetation of Oregon and Washington. Oregon State University Press, Corvallis.

Gaston, A. J. and I. L. Jones. 1998. The Auks: Alcidae (Bird families of the world series). Oxford University Press, Oxford.

Jodice, P. G. R. 1999. Behavioral ecology of Marbled Murrelets in forest and marine ecosystems of Oregon. Unpublished Ph.D. Thesis, Oregon State University, Corvallis.

Jodice, P. G. R. and M. W. Collopy. 2000. Activity patterns of Marbled Murrelets in Douglas-fir old-growth forests of the Oregon Coast Range. Condor 102: 272-285.

Kaiser, G. W., H. J. Barclay, A. E. Burger, D. Kangasniemi, D. J. Lindsay, W. T. Munro, W. R. Pollard, R. Redhead, J. Rice and D. Seip. 1994. National recovery plan for the Marbled Murrelet. Report No. 8. Ottawa, Canada: Recovery of nationally endangered wildlife committee, 36 pp.

Miller, S. L. and C. J. Ralph. 1995. Relationship of Marbled Murrelets with habitat characteristics at inland sites in California. Pages 205-215 *in* Ecology and conservation of the Marbled Murrelet (C. J. Ralph, Jr., G. L. Hunt, M. G. Raphael and J. F. Piatt, Eds.). Pacific Southwest Research Station Forest Service, USDA General Technical Report PSW-GTR-152. Albany.

Nelson. S. K. 1997. Marbled Murrelet (*Brachyramphus marmoratus*). The Birds of North America (A. Poole and F. Gill, Eds.). The Academy of Natural Sciences, Philadelphia, PA, and The American Ornithologists' Union, Washington, D.C.

O'Donnell, B. P., N. L. Naslund and C. J. Ralph. 1995. Patterns of seasonal variation in activity of Marbled Murrelets in forested stands. Pages 117-128 *in* Ecology and conservation of the Marbled Murrelet (C. J. Ralph, Jr., G. L. Hunt, M. G. Raphael and J. F. Piatt, Eds.). Pacific Southwest Research Station Forest Service, USDA General Technical Report PSW-GTR-152. Albany.

Paton, P. W. C. 1995. Marbled Murrelet inland patterns of activity: defining detections and behavior. Pages 113-116 *in* Ecology and conservation of the Marbled Murrelet (C.J. Ralph, Jr., G. L. Hunt, M. G. Raphael and J. F. Piatt, Eds.). Pacific Southwest Research Station Forest Service, USDA General Technical Report PSW-GTR-152. Albany.

Ralph, C. J., G. L. Hunt, Jr., M. G. Raphael and J. F. Piatt. 1995. Ecology and conservation of the Marbled Murrelet in North America: An overview. Pages 3-22 *in* Ecology and conservation of the Marbled Murrelet (C. J. Ralph, Jr., G. L. Hunt, M. G. Raphael and J. F. Piatt, Eds.). Pacific Southwest Research Station Forest Service, USDA General Technical Report PSW-GTR-152. Albany.

Ralph, C. J., S. K. Nelson, M. M. Shaughnessy and S. M. Miller. 1994. Methods for surveying Marbled Murrelets in forests. Pacific Seabird Group Technical Paper No. 1: 27 pp. Seattle.

Ralph, C. J. and S. L. Miller. 1995. Offshore population estimates of Marbled Murrelets in California. Pages 353-360 *in* Ecology and conservation of the Marbled Murrelet (C. J. Ralph, Jr., G. L. Hunt, M. G. Raphael and J. F. Piatt, Eds.). Pacific Southwest Research Station Forest Service, USDA General Technical Report PSW-GTR-152. Albany.

Ribic, C. A. and L. M. Ganio. 1996. Power analysis for beach surveys of marine debris. Marine Pollution Bulletin 32:554-557.

Rodway, M. S., H. M. Regehr and J-P. L. Savard. 1993. Activity levels of Marbled Murrelets in different inland habitats in the Queen Charlotte Islands, British Columbia. Canadian Journal of Zoology 71: 977-984.

SAS Institute, Inc. 1985. SAS procedures guide for personal computers, version 6 edition. SAS Institute Inc., Cary.

Schwagmeyer, P. L. and D. W. Mock. 1997. How to minimize sample sizes while preserving statistical power. Animal Behaviour 54:470-474.

Simon, J. 1995. Resampling Stats user's guide. Resampling Stats, Inc., Arlington.

Stauffer, H. B., C. J. Ralph and S. M. Miller. 1999. Incorporating detectability estimates into power and sample size selection to presence/absence surveys with application to Marbled Murrelets in California [abstract only]. Predicting species occurrences: issues of scale and accuracy.

Steidl, R. J., J. P. Hayes and E. Schauber. 1997. Statistical power analysis in wildlife research. Journal of Wildlife Management 61:270-279.

U.S. Fish and Wildlife Service. 1997. Recovery plan for the threatened Marbled Murrelet in Washington, Oregon, and California. 203 pp. Portland.

Varoujean, D. H. and W. A. Williams. 1995. Abundance and distribution of Marbled Murrelets in Oregon and Washington based on aerial surveys. *In* Ecology and conservation of the Marbled Murrelet. Pages 327-338 *in* Ecology and conservation of the Marbled Murrelet (C. J. Ralph, Jr., G. L. Hunt, M. G. Raphael and J. F. Piatt, Eds.). Pacific Southwest Research Station Forest Service, USDA General Technical Report PSW-GTR-152. Albany.